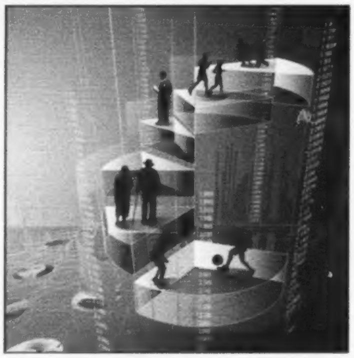
Component of Statistics Canada Catalogue no. 82-003-X
Health Reports

Article

Modelling risk factor information for linked census data: The case of smoking



by Claudia Sanmartin, Philippe Finès, Saeeda Khan, Paul Peters, Michael Tjepkema, Julie Bernier and Rick Burnett

June, 2013

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at infostats@statcan.gc.ca,

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

• Statistical Information Service	1-800-263-1136
• National telecommunications device for the hearing impaired	1-800-363-7629
• Fax line	1-877-287-4369

Depository Services Program

• Inquiries line	1-800-635-7943
• Fax line	1-800-565-7757

To access this product

This product, Catalogue no. 82-003-X, is available free in electronic format. To obtain a single issue, visit our website, www.statcan.gc.ca, and browse by "Key resource" > "Publications."

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "About us" > "The agency" > "Providing services to Canadians."

Published by authority of the Minister responsible for Statistics Canada

Cette publication est aussi disponible en français.

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard symbols

The following symbols are used in Statistics Canada publications:

.	not available for any reference period
..	not available for a specific reference period
...	not applicable
0	true zero or a value rounded to zero
0^s	value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p}	preliminary
^{r}	revised
x	suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E}	use with caution
F	too unreliable to be published
*	significantly different from reference category ($p < 0.05$)

Modelling risk factor information for linked census data: The case of smoking

by Claudia Sanmartin, Philippe Finès, Saeeda Khan, Paul Peters, Michael Tjepkema, Julie Bernier and Rick Burnett

Abstract

Background

Statistics Canada has initiated a series of data linkages of Census of Population long form and health outcome data. These linked data lack risk factor information. This study assesses the feasibility of using statistical modelling techniques to assign smoking status to census respondents.

Data and methods

The 2000/2001 Canadian Community Health Survey (CCHS) was used to develop age-/sex-specific predictive models to model smoking status based on variables available on the 1991 Census. The 2002/2003 CCHS was used to validate the modelled variable. Data from the 2002/2003 CCHS linked to data from the Hospital Morbidity Database (2001/2002 to 2004/2005) were used to evaluate the use of modelled versus self-reported smoking status on smoking-related hospitalizations.

Results

For the current daily smoker models, income, education, marital status, dwelling ownership and region of birth were significant predictors. For the never smoker models, marital status, dwelling ownership, Aboriginal identity and region of birth were significant predictors. Modelled current daily smoker status was associated with increased odds of smoking-related hospitalization, compared with being a never smoker, even when adjusting for covariates.

Interpretation

This study demonstrates the feasibility of using statistical modelling techniques to assign smoking status to census data, provided socio-economic and identity information is available.

Keywords

Health surveys, hospitalization, ROC Curve, socio-economic factors, statistical models

Authors

Claudia Sanmartin (1-613-951-6059; Claudia.sanmarting@statcan.gc.ca), Philippe Finès (1-514-283-6847; Philippe.Fines@statcan.gc.ca), Saeeda Khan, Paul Peters, Michael Tjepkema, and Julie Bernier are with the Health Analysis Division at Statistics Canada, Ottawa, Ontario, K1A 0T6. Rick Burnett is with the Healthy Environment and Consumer Safety Branch at Health Canada.

Administrative data are increasingly used to monitor the health of the population and to better understand health service use and outcomes. Advantages of using administrative data for health research include large population-based cohorts, low collection costs, and reduced bias from loss to follow-up.[1-3] Despite these advantages, administrative data have limited individual-level information, frequently restricted to demographics such as age and sex, and often do not include socio-economic or risk factor information, which limits a broader understanding of health outcomes.

To overcome these deficiencies, ecological approaches have "appended" area-level measures, such as neighbourhood indicators of socio-economic status, to administrative data.[4-6] However, ecological methods are prone to potential misclassification, underestimation of effect sizes, and inability to adjust for competing factors.[7-9] Moreover, the results of area-based studies reflect the characteristics not only of the population, but also of the physical and social setting of the particular geographic regions.[10]

Statistical techniques have also been employed to indirectly adjust for missing data that are associated with health outcomes. For instance, partitioned regression uses information from ancillary sources to adjust for missing risk factors.[11] This approach depends on the availability of such information from external data sources or in the literature.

Increasingly, data linkage is being used to fill information gaps in administrative data. For example, individual-level information collected in health surveys has been linked to hospital records to study broad determinants of hospital utilization.[12-14] These linked data are rich in individual-level information, but sample size and coverage issues often restrict analyses of subgroups and less common outcomes.

To address this shortcoming, Statistics Canada initiated a series of projects to link information from the Census

of Population long form with health outcome information, namely, mortality, hospitalization and cancer.[15,16] These linked datasets offer extensive, individual-level socio-economic information and large sample sizes, but they lack information on risk factors such as smoking and obesity.

This study assesses the feasibility of using statistical modelling techniques to fill information gaps related to risk factors, specifically, smoking, in linked census data.[15] Based on the Canadian Community Health Survey (CCHS), predictive algorithms were developed to model smoking status using variables common to the CCHS and the 1991 long-form Census. The resultant smoking variable was validated by comparing the performance of modelled versus self-reported smoking status in predicting smoking-related hospitalizations based on linked health survey and hospital data. This was considered an important step, since understanding how the modelled information performs in analysis is critical to assessing the utility of this approach.

Methods

Data source

Data from the CCHS were used to develop and validate predictive models for smoking status. The CCHS is a cross-sectional survey providing information about the health, lifestyle and health care use of the non-institutionalized household population aged 12 or older in the provinces and territories. The survey excludes full-time members of the Canadian Forces and residents of Indian reserves and some remote areas. A detailed description of the CCHS is available elsewhere.[17]

The 2000/2001 CCHS, the cycle closest in time to the linked 1991 Census cohort, was used to construct the predictive models. The response rate was 85%, for a total sample of 131,535. The sample for the present study was restricted to respondents aged 25 or older, the age criterion applied to the linked 1991 Census cohort. Records with missing information on smoking status were excluded, resulting in a final CCHS sample of 104,204.

Data from the 2002/2003 CCHS were used to externally validate the predictive models. The response rate was 81%, for a total sample of 134,072. Similar exclusions resulted in a final validation sample of 107,398.

The 2002/2003 CCHS data linked to the Hospital Morbidity Database (HMDB) (2001/2002 to 2004/2005) were used to evaluate associations between the modelled versus the self-reported smoking variable and smoking-related hospitalizations. The HMDB is a person-level administrative dataset representing inpatient hospitalizations from most acute care hospitals and some psychiatric, chronic and rehabilitation hospitals in Canada.[18] Data linkage was conducted among CCHS respondents living outside Quebec who agreed to link and who provided a valid personal health number (n=81,364). Similar exclusions were applied to the linked data (age 25 or older; missing smoking data), yielding a final sample of 52,396. Details about the data linkage are provided elsewhere.[12,19]

Development of predictive models

Separate models were constructed to predict two smoking categories: *current daily smokers* and *never smokers*. Smoking categories were derived based on self-reported information in the CCHS.[20] Current daily smokers were defined as respondents who reported that they smoked on a daily basis (1=yes, 0=no). Never smokers were those who reported that they had never smoked or had smoked fewer than 100 cigarettes in their lifetime (1=yes, 0=no). Attempts to predict former smokers were unsuccessful, as models were unable to discriminate between current, never and former smokers.

To be used to predict smoking status, CCHS variables had to be available in the census (long form) and to have been shown to be or hypothesized to be associated with smoking. When possible, the CCHS variables were coded to match the census variable definitions. Economic, socio-demographic, housing and ethno-cultural variables were used to predict smoking status (Table 1).

Multivariate logistic regression models were constructed to predict the probability of being a current daily smoker and a never smoker. Age-/Sex-specific models were developed because preliminary analyses revealed variability in the factors associated with smoking status across age and sex groups. The full study sample was used in both sets of models so that each CCHS respondent had a probability estimate of being a current daily smoker and a probability estimate of being a never smoker. The stepwise technique was applied to ensure selection of a parsimonious list of variables for each age and sex group; variables were included in the model by decreasing strength of significance. Survey weights were used, and the bootstrap technique was applied to the final multivariate regression models to adjust for the complex design of the CCHS. The models were developed using SAS's PROC LOGISTIC version 9.1.

Model-specific thresholds were established to classify respondents into smoking categories. Specifically, Receiver Operating Characteristic (ROC)[21-24] Curves were generated to determine age-/sex-specific optimal probability thresholds. If the estimated probabilities of being a current daily smoker or a never smoker exceeded the optimal thresholds, individuals were identified as positive cases. Optimal thresholds were generated to balance between false positives and negatives, with the aim of reducing the former. Given the large sample sizes associated with the census, focussing on true positives provides a more accurate model, even if a large number of false negatives are generated.

Model validation was assessed based on Area Under the Receiver Operating Curve (AUC), which is a plot representing a plot of sensitivity versus 1 minus specificity. In addition, the percentage of cases accurately predicted was calculated by comparing smoking status based on self-reported and modelled information.

Assignment of smoking status

The predicted probabilities for current daily smoker and for never smoker were used to assign each individual to both a current daily smoker category and a never smoker category, based on the sex-/age-specific thresholds. Both classification systems were then used to make a final assignment:

Never smoker	Current daily smoker: Yes	Current daily smoker: No
Yes	Unclassifiable	Never smoker
No	Current daily smoker	Other

For example, respondents whose probability of being never smokers exceeded the age-/sex-specific threshold, and whose probability of being current daily smokers was below the age-/sex-specific threshold were classified as never smokers. Respondents identified as being both current daily smokers and never smokers were deemed *unclassifiable* and removed from further analysis. Respondents classified as *other* were determined to be neither current daily smokers nor never smokers; they could be occasional smokers or former smokers, or might represent false negatives.

Additional threshold bands, defined as optimal threshold +/-0.05 or +/- 0.10, were generated to conduct sensitivity analyses. If the predicted probabilities were greater (lesser) than the upper (lower) threshold, respondents were identified as positive (negative) cases with respect to current daily smoker and never smoker status. This was deemed appropriate because the predicted value of the outcome was not the end product of the analysis, but rather, the appropriate assignment of smoking status.

Application of modelled smoking status

Linked 2002/2003 CCHS and 2001/2002 to 2004/2005 hospital data were used to determine how the modelled smoking variable performed in analyses of health outcomes. The objectives were: 1) to compare the association between smoking status and smoking-related hospitalizations using modelled versus self-reported smoking status; and 2) to assess the effect of using modelled smoking status on covariates also used to predict smoking status (for example, income, education). It was hypothesized that the effect size of the covariates may be reduced when using modelled smoking status, since similar variables were also used to predict smoking status.

A two-year follow-up period from the time individuals responded to the CCHS was examined to identify those who had at least one smoking-related hospitalization, defined as respiratory disease, cardiovascular or cancer-related admissions (based on ICD-9/10 and ICD-10-CA codes) reported as the primary diagnosis.[25] Logistic regression analyses were conducted to compare the results of using modelled versus self-reported smoking status: never smoker (reference group), current daily smoker, and other. A model-building approach was used to generate unadjusted models (Model 1: smoking status only), partially adjusted models (Model 2: smoking status + age and sex), and fully adjusted models (Model 3: Model 2 + additional socio-economic variables).

Survey weights for the linked CCHS file were adjusted by Statistics Canada to control for non-response to the survey and for the exclusion of records of respondents who did not agree to link and/or did not provide plausible health numbers. The bootstrap technique was applied to all analyses to account for the complex survey design in the estimate of variance and confidence intervals.

Results

Study population

Based on responses to the 2000/2001 CCHS, approximately 41% of the household population aged 25 or older were never smokers, and 26% were current daily smokers (Table 1). The majority of people were married or in a common-law relationship (71%), were employed (64%), owned their dwelling (73%), lived with at least one other person (85%), and had been born in Canada (76%). Around 40% had at least some postsecondary

Table 1
Percentage distribution (weighted) of selected characteristics of study sample, household population aged 25 or older, Canada, 2000/2001

Selected characteristics	Percent	95% confidence interval from	95% confidence interval to
Smoking status			
Never smoker	41.4	41.0	41.9
Current daily smoker	26.1	25.7	26.5
Other	32.5	32.0	32.9
Sex			
Women	51.2	51.1	51.4
Men	48.8	48.6	48.9
Age group			
25 to 44	46.4	46.3	46.6
45 to 64	35.7	35.6	35.8
65 or older	17.9	17.8	17.9
Marital status			
Single, never married	13.7	13.5	14.1
Married/Common-law	71.0	70.4	71.3
Separated	3.0	2.8	3.1
Divorced	5.7	5.6	6.0
Widowed	6.6	6.4	6.7
Education			
Less than secondary graduation	23.4	23.0	23.8
Secondary graduation	37.9	37.4	38.3
A least some postsecondary	19.8	19.4	20.2
University degree	19.0	18.5	19.4
Income quintile			
Lowest	17.4	16.9	17.7
Lower-middle	17.9	17.5	18.2
Middle	18.1	17.8	18.5
Upper-middle	18.1	17.7	18.4
Highest	18.9	18.5	19.3
Missing	9.7	9.5	10.0
Employment status			
Employed	63.9	63.6	64.3
Not in labour force	32.6	32.3	33.0
Unemployed	3.4	3.3	3.6
Dwelling ownership			
No	27.5	27.1	28.1
Yes	72.5	71.9	72.9
Household size			
One	14.8	14.6	15.2
Two	34.8	34.5	35.3
Three	18.8	18.4	19.2
Four	19.4	19.0	19.8
Five	8.0	7.7	8.2
Six or more	4.2	3.9	4.4
Aboriginal ancestry			
No	97.2	97.1	97.4
Yes	2.8	2.6	2.9
Visible minority status			
Non-White	14.1	13.6	14.5
White	85.9	85.5	86.4
Region of birth			
Canada	76.2	75.9	76.8
Other North America	1.3	1.2	1.4
South/Central America/Caribbean	2.5	2.3	2.7
Europe and Oceania	11.0	10.7	11.3
Africa	1.2	1.1	1.4
Asia	7.7	7.4	8.0
Rural/Urban indicator			
Rural area (farm and non-farm)	18.3	17.7	18.8
Small urban area (less than 30,000)	15.0	14.7	15.5
Urban area (30,000 to 99,999)	9.7	9.3	10.0
Urban area (100,000 to 499,999)	11.4	11.1	11.7
Urban area (500,000 or more)	45.6	45.0	46.1

Source: 2000/2001 Canadian Community Health Survey.

education. Just under half (46%) lived in urban areas with more than 500,000 inhabitants.

Predictive models

The variables that were important in predicting smoking status differed by age group and sex and are presented in order of significance (Table 2). For models predicting current daily smoker, income quintile, education, marital status, dwelling ownership and world region of birth were significant predictors across all age and sex groups. For the never smoker models, marital status, dwelling ownership, Aboriginal ancestry and world region of birth were significant predictors across all age and sex groups. When the age-/sex-specific optimal thresholds were applied to the probabilities generated from the predictive models, close to 80% of respondents were assigned to either the current daily smoker or never smoker categories, 7.0% were *unclassifiable*, and 14.6% were classified as *other*.

AUC values ranged from 0.59 to 0.73 for the current daily smoker models, and from 0.60 to 0.70 for the never smoker models. Using optimal thresholds, the percentage of cases correctly predicted based on modelled values ranged from 54% to 67% for current daily smoker, and from 57% to 65% for never smoker, with AUC values decreasing with advancing age. The percentage of correctly predicted cases decreased when the wider threshold bands (optimal +/- 0.05 and optimal +/- 0.10) were used.

Modelled versus self-reported smoking status

Logistic models were developed to compare the performance of modelled versus self-reported smoking status in predicting smoking-related hospitalizations, and to assess the effect of using the modelled variable on covariates that had

Table 2
Variables predicting smoking status (in order of significance), by sex and age group, household population aged 25 or older, Canada, 2000/2001

Smoking status model	Men			Women		
	25 to 44	45 to 64	65 or older	25 to 44	45 to 64	65 or older
Curent daily smoker						
Explanatory variables	Education Home ownership Region of birth Marital status Employment status Income quintile Aboriginal ancestry Household size	Home ownership Education Region of birth Marital status Income quintile Number of bedrooms Employment status Aboriginal ancestry Household size Rural/Urban	Marital status Income quintile Household size Region of birth Home ownership Education	Education Region of birth Home ownership Marital status Income quintile Household size Aboriginal ancestry Employment status Ethnicity	Region of birth Home ownership Education Marital status Income quintile Household size Number of bedrooms Aboriginal ancestry Ethnicity	Home ownership Ethnicity Education Marital status Region of birth Income quintile Household size
AUC	0.702	0.677	0.643	0.729	0.672	0.586
Optimal threshold	0.305	0.257	0.108	0.269	0.223	0.105
% correct predictions (optimal threshold)	63.5%	62.2%	59.9%	66.5%	61.9%	54.4%
% correct predictions (optimal threshold +/- 0.05)	46.0%	42.1%	12.5%	50.8%	39.9%	12.8%
% correct predictions (optimal threshold +/- 0.10)	37.2%	26.1%	2.8%	38.6%	22.7%	1.1%
Never smoker						
Explanatory variables	Education Region of birth Home ownership Employment status Marital status Number of bedrooms Aboriginal ancestry Ethnicity Household size Income quintile	Education Region of birth Home ownership Household size Marital status Employment status Number of bedrooms Income quintile Aboriginal ancestry Rural/Urban Ethnicity	Region of birth Education Employment status Home ownership Marital status Aboriginal ancestry Ethnicity Household size Rural/Urban	Region of birth Education Home ownership Household size Marital status Aboriginal ancestry Ethnicity Income quintile Employment status	Region of birth Home ownership Marital status Education Number of bedrooms Aboriginal ancestry Ethnicity Income quintile Household size	Region of birth Marital status Home ownership Ethnicity Aboriginal ancestry Income quintile Household size Rural/Urban Number of bedrooms
AUC	0.679	0.659	0.596	0.703	0.637	0.591
Optimal threshold	0.574	0.698	0.731	0.552	0.549	0.385
% correct predictions (optimal threshold)	62.3%	60.9%	57.3%	64.6%	59.4%	56.3%
% correct predictions (optimal threshold +/- 0.05)	46.4%	41.6%	21.7%	50.2%	36.0%	28.0%
% correct predictions (optimal threshold +/- 0.10)	38.3%	27.4%	10.3%	37.8%	24.7%	12.0%

AUC= Area Under Receiver-Operating Curve

also been used to predict smoking status (for example, income, education).

As expected, based on *self-reported* smoking status, being a current daily smoker rather than a never smoker was associated with increased odds of at least one smoking-related hospitalization in both unadjusted and adjusted models (Table 3). The association was similar, but weaker, when *modelled* smoking status was used. Unadjusted odds ratios for *modelled* current daily smoker status ranged from 1.81 to 2.99 across various threshold definitions. The odds ratios remained significant in the fully adjusted models using the optimal threshold (OR: 1.30) and the optimal threshold +/- 0.05 (OR: 1.52), but were lower than the odds when self-reported smoking status was used (OR: 2.19).

Overall, variables significantly associated with smoking-related hospitalizations in the model using self-reported smoking status (Model A) remained significant when modelled smoking status was used instead (Table 4). Older age, Aboriginal identity, widowhood, lower education and being unemployed or not in the labour force were consistently associated with higher odds of a smoking-related hospitalization. Being female and being never married were associated with lower odds of a smoking-related hospitalization. Income was not associated with smoking-related hospitalizations, regardless of whether the model incorporated self-reported or modelled smoking status.

Discussion

This study examined the feasibility of using statistical modelling techniques to predict smoking status, and then assessed the association between the modelled variable and smoking-related hospitalizations. The set of socio-economic and demographic characteristics that were predictive of smoking status varied by age and sex, which highlights the importance of developing age-/sex-specific models.

Model validation revealed AUC values close to 0.70 for most of the age/sex models using the optimal threshold, somewhat below values achieved in other studies.[26] However, this project is unique in that no health-related variables were used to predict smoking status, whereas in other studies, health-related characteristics are commonly used to predict outcomes such as hospitalization and mortality. AUC values were consistently low for the female aged 65 or older models for both current daily smokers and never smokers. The ability of the predictive models to accurately assign smoking status decreased when threshold levels were relaxed.

Table 3
Unadjusted and adjusted odds ratios relating self-reported and modelled smoking status to smoking-related acute care hospitalizations, household population aged 25 or older, Canada excluding Quebec, 2002/2003 to 2004/2005

	Current daily smoker		
	Odds ratio	95% confidence interval from	95% confidence interval to
Self-reported smoking status			
Unadjusted (smoking status only)	1.71*	1.38	2.11
Partially adjusted (age and sex)	2.41*	1.92	3.02
Fully adjusted†	2.19*	1.74	2.77
Modelled smoking status			
Optimal threshold			
Unadjusted (smoking status only)	1.81*	1.53	2.12
Partially adjusted (age and sex)	1.97*	1.65	2.34
Fully adjusted†	1.30*	1.04	1.63
Optimal threshold +/- 0.05			
Unadjusted (smoking status only)	2.53*	1.87	3.43
Partially adjusted (age and sex)	2.69*	1.96	3.67
Fully adjusted†	1.52*	1.02	2.26
Optimal threshold +/- 0.10			
Unadjusted (smoking status only)	2.99*	1.89	4.73
Partially adjusted (age and sex)	2.72*	1.70	4.36
Fully adjusted†	1.07	0.61	1.85

* significantly different from "never smoker" ($p<0.05$)
† age, sex, Aboriginal ancestry, visible minority, marital status, education, income, employment
Source: 2002/2003 Canadian Community Health Survey; 2001/2002 to 2004/2005 Hospital Morbidity Database

What is already known on this subject?

- Increasingly, administrative data are being used to monitor the health of the population and understand health service use and outcomes.
- Data linkage has been used to fill information gaps in administrative data.
- In some linked data, information gaps remain (for example, risk factor information).

What does this study add?

- This study assesses the feasibility of using statistical modelling techniques to fill information gaps, specifically smoking status, in linked long-form census data.
- Predictive algorithms to model smoking status were developed, based on the Canadian Community Health Survey.
- Regression analysis demonstrated the viability of using the modelled smoking variable to examine associations between smoking status and smoking-related hospitalizations.

This study was motivated by the need to provide risk factor information in census data that are linked to administrative records to study characteristics associated with health outcomes. Hence, it was critical to demonstrate the feasibility of using modelled smoking status in a research context. The linked survey and hospital data offered this opportunity.

The results of the regression analysis that compared associations between modelled versus self-reported smoking status and smoking-related hospitalizations demonstrated the viability of the modelled variable. Modelled smoking status behaved like self-reported smoking status in terms of direction of association and significance, albeit with smaller effect sizes. Furthermore, the use of modelled smoking status did not eliminate associations between hospital-

Table 4
Adjusted odds ratios relating self-reported and modelled smoking status and selected characteristics to smoking-related acute care hospitalizations, household population aged 25 or older, Canada excluding Quebec, 2002/2003 to 2004/2005

	Model A (Self-reported smoking)			Model B (Modelled smoking - optimal threshold)			Model C (Modelled smoking - optimal threshold +/- 0.05)			Model D (Modelled smoking - optimal threshold +/- 0.10)		
	Adjusted odds ratio	95% confidence interval		Adjusted odds ratio	95% confidence interval		Adjusted odds ratio	95% confidence interval		Adjusted odds ratio	95% confidence interval	
Selected characteristics		from	to		from	to		from	to		from	to
Smoking status												
Never smoker†	1.00	...	...	1.00	...	...	1.00	...	...	1.00	...	...
Current daily smoker	2.19*	1.74	2.77	1.30*	1.04	1.63	1.52*	1.02	2.26	1.07	0.61	1.85
Other smoker	1.51*	1.28	1.79	1.21	0.95	1.53	0.88	0.35	2.17	0.99	0.55	1.80
Sex												
Men†	1.00	...	...	1.00	...	...	1.00	...	...	1.00	...	...
Women	0.51*	0.44	0.60	0.47*	0.40	0.54	0.47*	0.40	0.55	0.46*	0.39	0.54
Age group												
25 to 44†	1.00	...	...	1.00	...	...	1.00	...	...	1.00	...	...
45 to 64	4.72*	3.18	7.02	4.77*	3.19	7.14	4.73*	3.14	7.14	4.65*	3.07	7.04
65 to 75	9.46*	6.43	13.91	8.94*	5.83	13.70	8.46*	5.35	13.36	8.38*	5.19	13.53
Marital status												
Married/Common-law†	1.00	...	...	1.00	...	...	1.00	...	...	1.00	...	...
Widowed	1.28*	1.06	1.55	1.29*	1.06	1.57	1.30*	1.07	1.58	1.30*	1.07	1.58
Separated/Divorced	1.05	0.77	1.41	1.05	0.76	1.47	1.07	0.75	1.54	1.16	0.79	1.71
Single, never married	0.62*	0.46	0.85	0.62*	0.46	0.85	0.62*	0.46	0.84	0.62*	0.46	0.85
Education												
Less than secondary graduation	1.83*	1.32	2.55	1.83*	1.29	2.61	1.72*	1.19	2.47	2.03*	1.47	2.80
Secondary graduation	1.49*	1.08	2.05	1.53*	1.10	2.12	1.41	1.00	1.99	1.63*	1.20	2.21
A least some postsecondary	1.42	0.93	2.15	1.44	0.94	2.19	1.31	0.86	2.01	1.52*	1.02	2.26
University degree†	1.00	...	...	1.00	...	...	1.00	...	...	1.00	...	...
Employment status												
Employed†	1.00	...	...	1.00	...	...	1.00	...	...	1.00	...	...
Unemployed	2.08*	1.63	2.66	2.10*	1.63	2.70	2.10*	1.64	2.70	2.14*	1.65	2.77
Not in labour force	5.32*	4.12	6.86	5.17*	3.94	6.80	5.20*	3.97	6.81	5.29*	4.02	6.96
Income quintile												
Lowest	1.02	0.74	1.40	1.00	0.72	1.39	0.99	0.72	1.36	1.05	0.76	1.45
Lower-middle	0.93	0.71	1.24	0.92	0.68	1.23	0.90	0.67	1.20	0.95	0.72	1.26
Middle	0.86	0.64	1.16	0.86	0.64	1.15	0.83	0.62	1.12	0.88	0.66	1.19
Upper-Middle	0.98	0.75	1.29	0.98	0.73	1.32	0.97	0.72	1.30	0.99	0.74	1.34
Highest†	1.00	...	...	1.00	...	...	1.00	...	...	1.00	...	...
Aboriginal ancestry												
Yes	2.85*	1.05	7.69	3.07*	1.11	8.51	2.97*	1.07	8.28	3.49*	1.27	9.58
No†	1.00	...	...	1.00	...	...	1.00	...	...	1.00	...	...
Visible minority												
Yes	0.68	0.45	1.03	0.67	0.44	1.02	0.72	0.46	1.13	0.62*	0.41	0.95
No†	1.00	...	...	1.00	...	...	1.00	...	...	1.00	...	...

† reference category
* significantly different from reference category ($p<0.05$)
... not applicable
Source: 2002/2003 Canadian Community Health Survey, 2001/2002 to 2004/2005 Hospital Morbidity Database.

ization and other covariates (for example, marital status, education, employment status). The association between modelled smoking status and hospitalization was reduced in the fully adjusted models, but remained significant.

Limitations

This study has several limitations. The CCHS excludes specific subgroups (Canadian Forces, residents of Indian reserves and some remote areas) and people who did not agree to link their data; these exclusions may have affected the final models used to predict smoking status. The feasibility of using modelled smoking status was assessed only in the context of smoking-related hospitalizations using logistic regression analysis. Further investigation is needed to determine if this modelled variable can be used in studies employing alternative techniques (for example, survival analysis) and/or outcomes (for example, mortality).

Conclusion

Data linkage is a cost-effective method of obtaining person-level data to study health outcomes at the population level. However, data gaps, specifically, a lack of risk factor information, may exist. This study demonstrates the feasibility of using statistical modelling techniques to implement information in data sources.

References

1. Roos LL, Nicol JP, Cageorge SM. Using administrative data for longitudinal research: comparisons with primary data collection. *Journal of Chronic Diseases* 1987; 40: 41-9.

2. Lezzoni LI. Using administrative data to study persons with disabilities. *The Milbank Quarterly* 2002; 80(2): 347-79.

3. Brackstone GJ. Issues in the use of administrative records for statistical purposes. *Survey Methodology* 1987; 13: 29-43.

4. Disano J, Goulet J, Muhajarine N, et al. Social-economic status and rates of hospital admission for chronic disease in urban Canada. *The Canadian Nurse* 2010; 106(1): 24-9.

5. Booth GL, Bisara P, Lipscombe LL, et al. Universal drug coverage and socioeconomic disparaties in major diabetes outcomes. *Diabetes Care* 2012; epub ahead of print.

6. Wilkins R, Berthelot JM, Ng E. Trends in mortality by neighbourhood income in urban Canada from 1971 to 1996. *Health Reports* 2002; 13: 1-28.

7. Hanley GE, Morgan S. On the validity of area-based income measures to proxy household income. *BMC Health Services Research* 2008; 8: 79.

8. Marra CA, Lynd LD, Harvard SS, Grubisic X. Agreement between aggregate and individual-level measures of income and education: a comparison across three patient groups. *BMC Health Services Research* 2011; 11: 69.

9. Southern DA, McLaren L, Hawe P, et al. Individual-level and neighbourhood income measures: agreement and association with outcomes in a cardiac disease cohort. *Medical Care* 2005; 43(11): 1116-22.

10. Pampalon R, Hamel D, Gamache P. A comparison of individual and area-based socio-economic data for monitoring social inequalities in health. *Health Reports* 2009; 20(4): 85-94.

11. Burnett RT, Cakmak S, Brion O, et al. An indirect adjustment method for multiple missing variables applicable to environmental epidemiology. *American Journal of Epidemiology*. Under review.

12. Rotermann M. Evaluation of the coverage of linked Canadian Community Health Survey and hospital inpatient records. *Health Reports* (Statistics Canada, Catalogue 82-003) 2009; 20(1): 45-51.

13. Wilkins K, Shields M, Rotermann M. Smokers' use of acute care hospitals: A prospective study. *Health Reports* 2009; 20(4): 75-83.

14. Sanmartin C, Khan S. *Hospitalizations for Ambulatory Care Sensitive Conditions (ACSC): The Factors That Matter.* Health Research Working Paper Series (Catalogue 82-622-X, No. 007) Ottawa: Statistics Canada, 2011.

15. Wilkins R, Tjepkema M, Mustard C, Choinière R. The Canadian census mortality follow-up study, 1991 through 2001. *Health Reports* 2008; 19(3): 25-43.

16. Peters PA, Tjepkema M. 1991-2011 Canadian census mortality and cancer follow-up study. *Proceedings of Statistics Canada Symposium 2010. Social Statistics: The Interplay among Censuses, Surveys and Administrative Data* (Catalogue. 11-522-XCB) Ottawa: Statistics Canada, 2011:150-6.

17. Béland Y, Dale V, Dufour J, Hamel M. The Canadian Community Health Survey: Building on the success from the past. *Proceedings of the American Statistical Association Joint Statistical Meetings 2005, Section on Survey Research Methods.* Minneapolis, Minnesota: American Statistical Association, 2005.

18. Statistics Canada, Household Survey Methods Division. *External Linkage of Person-oriented Information 1992/93 to 2000/01 Hospital Morbidity Files.* (unpublished). Ottawa: Statistics Canada, 2003.

19. Nadeau C. *Linking HPOI 1992-2005 to CCHS.* Household Suvey Methods Division internal documents. Ottawa: Statistics Canada, 2007.

20. Wong S, Shields M, Leatherdale S, et al. Assessment of validity of self-reported smoking status. *Health Reports* 2012; 23(1): 47-53.

21. Hanley JA, McNeil BJ. The meaning and use of the area under a Receiver Operating Characteristic (ROC) Curve. *Diagnostic Radiology* 1982; 143(1): 29-36.

22. Metz CE. Basic principles of ROC analysis. *Seminars in Nuclear Medicine* 1978; 8(4): 283-98.

23. Streiner DL, Cairney JC. What's under the ROC? An introduction to Receiver-Operating Characteristics Curves. *The Canadian Journal of Psychiatry* 2007; 52(2): 121-8.

24. Swets JA, Dawes RM, Monahan J. Psychological science can improve diagnostic decision. *Psychological Science in the Public Interest* 2000; 1(1): 1-26.

25. Balinuas D, Patra J, Rehm J, et al. Smoking-attributable morbidity: acute care hospital diagnoses and days of treatment in Canada, 2002. *BMC Public Health* 2007; 7: 247.

26. Rosella LC, Manuel DG, Burchill C, Stukel TA. A population-based risk algorithm for the development of diabetes: development and validation of the Diabetes Population Risk Tool (DPort). *Journal of Epidemiology and Community Health* 2011; 65(7): 613-20.